Anticipated Changes in Conducting Scientific Data-Analysis Research in the Big-Data Era

Kwo-Sen Kuo, Michael Seablom, Thomas Clune, and Rahul Ramachandran

A Big-Data environment is one that is capable of orchestrating quick-turnaround analyses involving large volumes of data for numerous simultaneous users. Based on our experiences with a prototype Big-Data analysis environment, we anticipate some important changes in research behaviors and processes while conducting scientific data-analysis research in the near future as such Big-Data environments become the mainstream. The first anticipated change will be the reduced effort and difficulty in most parts of the data management process. A Big-Data analysis environment is likely to house most of the data required for a particular research discipline along with appropriate analysis capabilities. This will reduce the need for researchers to download local copies of data. In turn, this also reduces the need for compute and storage procurement by individual researchers or groups, as well as associated maintenance and management afterwards.

It is almost certain that Big-Data environments will require a different "programming language" to fully exploit the latent potential. In addition, the process of extending the environment to provide new analysis capabilities will likely be more involved than, say, compiling a piece of new or revised code.We thus anticipate that researchers will require support from dedicated organizations associated with the environment that are composed of professional software engineers and data scientists. A major benefit will likely be that such extensions are of higher-quality and broader applicability than ad hoc changes by physical scientists.

Another anticipated significant change is improved collaboration among the researchers using the same environment. Since the environment is homogeneous within itself, many barriers to collaboration are minimized or eliminated. For example, data and analysis algorithms can be seamlessly shared, reused and re-purposed. In conclusion, we will be able to achieve a new level of scientific productivity in the Big-Data analysis environments.